

CJBB75 – 1 (G1310.00-11.30)

K. Osolsobě

Výuka: od 28. 2. 2022 každých 14 dní kontaktní, každých 14 dní úkol (viz harmonogram)

Podmínky ukončení: Průběžné plnění úkolů (5 odevzdaných referátů, závěrečný test).


Náplň dnešní hodiny 21. 2. 2022

Co je to korpus?

- Soubor textů
- elektronicky uložených a přístupných (korpusové manažery – programy, skrze něž lze ke korpusům přistupovat)
- má stanovený obsah (složený z textů záměrně vybraných dle zveřejněných kritérií)
- má stanovený rozsah/velikost (lze na něm pracovat s frekvenčními/kvantitativně měřitelnými údaji)
- obsahuje standardní anotace (metadata – údaje o každém textu a lingvistické interpretace, anotace jazykových jednotek – vnitřní anotace)

Registrace uživatele pro práci s ČNK (<http://ucnk.ff.cuni.cz/>)

---

Kdo jsme? 

---

Český národní korpus je **akademický projekt** založený v roce 1994 při **FF UK** a spravovaný **Ústavem Českého národního korpusu**. Jeho cílem je systematicky mapovat češtinu a další jazyky ve srovnání s ní. **Korpusy ČNK** jsou po **bezplatné registraci** otevřeny všem zájemcům o jazyk, kteří touží vědět, jak se čeština používá. [více...](#)

---

Korpusový manažer

Základy práce s korpusem přes Kontext

KonText

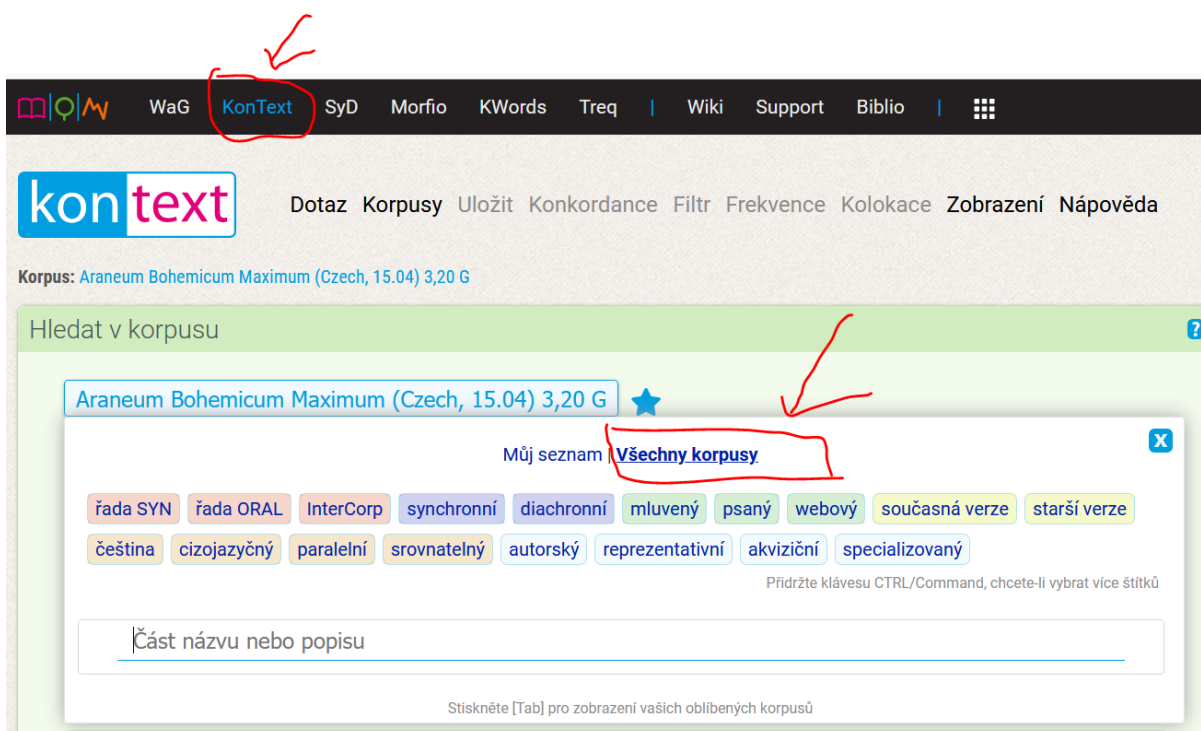


Dotaz

## Výběr korpusu

<https://wiki.korpus.cz/doku.php/cnk:uvod>

## V rozhraní KonText



## Jaké korpusy jsou k dispozici ?

Časové hledisko (synchronní / diachronní)

Hledisko textů (psané / mluvené, připravené/spontánní)

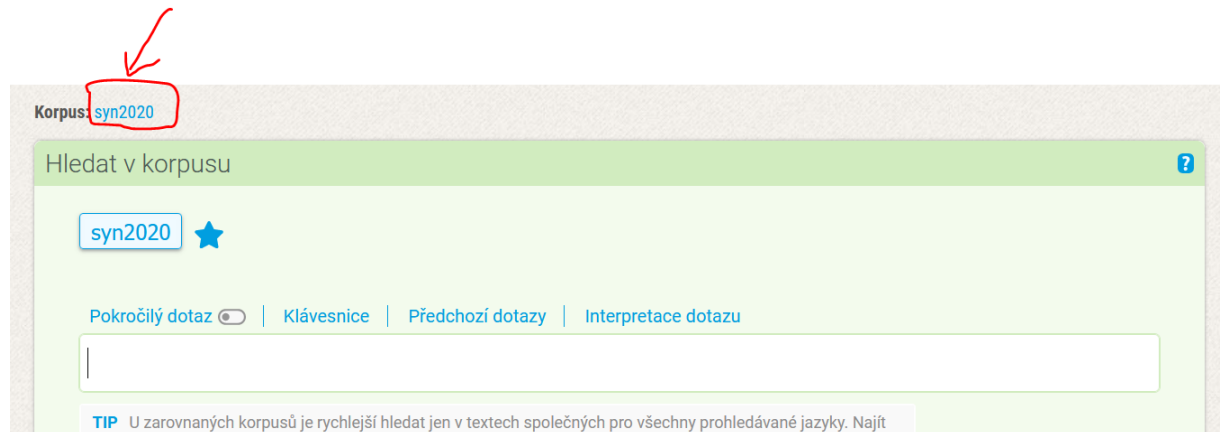
Hledisko žánru (vyvážené žánrově/ žánrově kompaktní – např. korpusy výhradně publicistické, nebo korpus soukromé korespondence, projekt Korpus českého verše).

Hledisko autora (autoři jsou rodilí mluvčí/ autoři se učí jazyk, v němž jsou texty vytvořeny jako tzv. druhý jazyk – learner corpora/žakovské korpusy, autorské korpusy založené na díle/korespondenci významných osobností).

Hledisko jazyka (jednojazyčné – např. čeština/ vícejazyčné, srovnatelné, paralelní).

Vícejazyčné paralelní korpusy – stejné texty – originál+překlad – zarovnání/alignment = jednotky, které si odpovídají, jsou propojeny / srovnatelné korpusy – různojazyčné i stejného jazyka vybudované stejným způsobem, mající stejné složení).

## Jak čteme informace o zvoleném korpusu?



The screenshot shows the top part of the corpus search interface. At the top left, the text 'Korpus: syn2020' is displayed, with 'syn2020' highlighted by a red box and a red arrow pointing to it. Below this is a search bar with the text 'Hledat v korpusu'. Inside the search bar, 'syn2020' is entered and marked with a blue star. Below the search bar, there are links for 'Pokročilý dotaz', 'Klávesnice', 'Předchozí dotazy', and 'Interpretace dotazu'. At the bottom of the search bar, there is a tip: 'TIP U zarovnaných korpusů je rychlejší hledat jen v textech společných pro všechny prohledávané jazyky. Najít'.

## Stručné info.



The screenshot shows the detailed information for the selected corpus 'syn2020'. The title 'syn2020' is prominently displayed. Below it, the following information is provided: 'Popis:' (empty), 'Velikost: 121 826 797 pozic' (with a red arrow pointing to the number), 'Webová stránka: <http://wiki.korpus.cz/doku.php/cnk:syn2020>' (with a red arrow pointing to the URL), and 'Štítky: čeština, reprezentativní, synchronní, řada SYN, psaný' (with a red bracket underlining the tags). The search bar on the left is also visible, showing 'syn2020' and a blue star.

## Citace:



The screenshot shows the citation information for the corpus 'syn2020'. The title 'Citační informace:' is displayed. Below it, the following information is provided: 'Korpus jako zdroj dat:' (empty), 'Křen, M. – Cvrček, V. – Henyš, J. – Hnátková, M. – Jelínek, T. – Koček, J. – Kovářiková, D. – Křivan, J. – Milička, J. – Petkevič, V. – Procházka, P. – Skoumalová, H. – Šindlerová, J. – Škrabal, M.: SYN2020: reprezentativní korpus psané češtiny. Ústav Českého národního korpusu FF UK, Praha 2020. Dostupný z WWW: <http://www.korpus.cz>', 'Obecné odkazy:' (empty), and 'http://www.korpus.cz'.

## Proč je třeba citovat korpusy?

Korpusy ÚČNK vznikly jako výsledek státní podpory GAČR. V korpusech jsou texty, které mnohdy spadají pod autorská práva. ÚČNK poskytl smluvní garance „poskytovatelům textů“.

## Jak číst informace o velikosti korpusu:


Termíny: viz <http://wiki.korpus.cz/doku.php>

[http://wiki.korpus.cz/doku.php/pojmy:prehled\\_pojmu](http://wiki.korpus.cz/doku.php/pojmy:prehled_pojmu)

## Tokenizace

Token je nejmenší jednotka textu, většinou se jedná o grafické slovo (tj. řetězec alfabetských znaků oddělený mezerou v textu), resp. o jednu jeho konkrétní realizaci. V některých případech je jedno grafické slovo rozděleno na dvě (např. *mohu -li*), často je také z praktických důvodů (pro snadné vyhledávání) oddělována interpunkce od předcházejícího slova (3 tokeny: *řekl , že*). O jednotlivých tokenech v korpusu se také mluví jako o pozicích.

Všimněme si:

A teď si vezměte , že <b>se</b> mluvčího do Prahy	<b>ptám</b>	na maličkost , kterou by mi řekl kdejaký úředník .
- Podle policistů z šumperského dopravního inspektorátu <b>se</b> lidé nejčastěji	<b>ptají</b>	na používání bezpečnostních pásů při přepravě dětí . " Je
to , proč se tak doposud nestalo , <b>se</b> nás	<b>ptal</b>	i místostarosta Pavelec , " vyjádřila se Kelárková s tím
Skalice u České Lípy – Když <b>se</b> ho televizní reportér	<b>ptal</b>	, jaká by měla být jeho vyvolená , odpověděl naprosto
koupit už v pátek . " Lidé <b>se</b> po něm	<b>ptali</b>	. Prodáváme ho s předstihem , protože v neděli máme
, a chce slyšet můj názor . Včera <b>se</b> mě	<b>ptal</b>	třeba na to , jestli nevím , co by se
k nám nastěhovat . První , na co <b>se</b> totiž	<b>ptají</b>	, je , jestli je v Zádveřicích škola . Také
projekty pro děti . Již během roku <b>se</b> mne lidé	<b>ptají</b>	, jaká překvapení zase připravíme . Moc bych chtěla poděkovat
tu nejsou všichni nejlepší , ale na to <b>se</b> historie	<b>neptá</b>	. « Sám říkal , že pokud nezíská medaili ,
pojistných podmínek , protože to je součást jejich školení .	<b>Ptali</b>	jsem <b>se</b> , co by se stalo , kdybychom to
nic jsem okolo sebe nevnímala , protože jsem <b>se</b> znovu	<b>ptala</b>	sama sebe , proč šel zase běhat . Tak jsem
bez onoho jmění neexistovaly . V nadpisu jsem <b>se</b> vás	<b>ptala</b>	dotazem , který na svých webových stránkách pokládají autoři výše
španělsky . „ Ty ses to někde učila ? “	<b>ptal</b>	jsem <b>se</b> . „ Ne , “ odpověděla , „
 lidé , kteří jsou v pivovaru na exkurzích , udiveně	<b>ptají</b>	, proč mají téměř prázdné sklady sudového a lahového piva
. Mimochodem ( tento dopis je plný mimochoďů ) ,	<b>ptal</b>	<b>ses</b> mne z Prahy , zda Chris stále cituje duchy

Pro rodilého mluvčího je zvrtné *se* součástí reflexiva tantum *ptát se*. Pro automatickou morfologickou analýzu jde o dva samostatné tokeny.

## Pozice

V souvislosti s tím, že každý text, který vstupuje do korpusu, prochází procesem tokenizace, se o jednotkách v korpusu nemluví jako o slovech, ale častěji jako o **pozicích**. Tokenizace se přitom u jednotlivých korpusů může lišit, pozicí se tak v různých korpusech může myslet různě vymezená jednotka.

## Lemmatizace a taggování

Lemma je reprezentativní slovníková podoba hesla, při automatickém zpracování jazyka je pak tato podoba v procesu lemmatizace přidělována každé formě v korpusu.

Přístupy k lemmatizaci se mohou v drobnostech lišit, obecně však platí, že

- lemma každého českého substantiva je jeho **nom. sg.** (tvary *lesům, lesy, lesích* mají lemma *les*)
- u adjektiv je to **nom. sg. mask. pozitiv** (tvary *chytrého, chytrou, chytrejma, nejchytrější* mají lemma *chytrý*)
- u zájmen je to **nom. sg. mask.** (tvary *ta, to, ti, tomu* mají lemma *ten*)
- u sloves je to **infinitiv** (tvary *chodil, chodíš, chodíme* mají lemma *chodit*)

Lemma jako jednotka vzniká abstrakcí morfologických vlastností [slovního tvaru](#) (označovaného jako word nebo forma), představuje tedy množinu forem se stejným kořenem lišící se pouze morfologickými afíxy, příp. pravopisnou variantou. V některých koncepcích se pak k lemmatu řadí i vybrané varianty slovtvorné.

Představme si následující dialog, která z variant je podle vás více na místě, A nebo B?

A

- No víš, viděl jsem takovou *fuškunkuli* a ona ti měla na hlavě takovou *kumušinku paškovanou* a ona si ji ještě *vygárovala*.
- Co je to *fuškunkuli* a *kumušinku paškovanou*? A co je to *vygárovala*?

B

- No víš, viděl jsem takovou *fuškunkuli* a ona ti měla na hlavě takovou *kumušinku paškovanou* a ona si ji ještě *vygárovala*.
- Co je to *fuškunkule* a *kumušinka paškovaná*? A co je to *vygárovat*?

Uvědomte si, že lemmatizace je činnost, kterou dnes automaticky provádí řada nástrojů od vyhledávačů na webu přes on-line slovníky. Jde ale o schopnost, kterou nabývá i dítě během akvizice jazyka, kterou má mluvčí, když se dotazuje na neznámé slovo, kterou aplikujeme, když hledáme v cizojazyčném slovníku (např. význam tvaru *went* nenajdeme ve slovníku angličtiny pod *w*, ale pod *g*).

## Desambiguace

Desambiguace (někdy též disambiguace, z lat. *dis-* vyjadřuje zápor, *ambo* oba, česky zjednoznačnění) je část (většinou automatického) procesu [anotace](#) jazykových dat, které vstupují do korpusu.

Zjednoznačněním se většinou myslí odstranění homonymie, čili jednoznačná interpretace slovního tvaru či skupiny slovních tvarů nebo věty na základě kontextu či mimojazykové situace. Desambiguace se obecně týká všech jazykových rovin, nejčastěji se ovšem v korpusech češtiny uplatňuje na rovině [morfologické](#) (zahrnující [lemmatizaci](#) a přiřazení náležitých morfologických údajů slovnímu tvaru na základě kontextu).

Např. ve větě *Větry vanou od západu.* se při morfologické interpretaci věty nejprve přiřadí morfologickou analýzou tvaru *vanou* dvě lemmata a dvě morfologické interpretace:

1. lemma = *vana*, subst. fem. sg. instr.
2. lemma = *vát*, 3. os. pl. přez,

a poté se při desambiguaci vybere náležitá 2. interpretace.

V následujících větách si všimněte, jak je třeba nejednoznačný tvar **sil**, který lze interpretovat jako a) genitiv plurálu feminina k lemmatu **сила**, b) genitiv plurálu neutra k lemmatu **сило**, c) variantní tvar l-ového přičestí maskulina singuláru slovesa **сít**.

a) *Podle jeho názorů je internet jednou ze **sil**, která dostala Ameriku na špici*

b) *Z jednoho ze **sil** začala náhle tryskat čpící tekutina a ocelová konstrukce jedné z věží se zhroutila.*

c) *Raná variační fantazie na lidový nápěv **Sil** jsem proso dala oběma protagonistům možnost ukázat jejich virtuozitu.*

Někdy může být situace dosti složitá:

*Odstupující ministr informatiky Vladimír Mlynář podle serveru iDNES odmítl nabídku premiéra Grosse **stát se** šéfem Českého telekomunikačního úřadu.*

*Potřeboval **stát se** svým zločineckým gangem.*

**Jaké přednosti má lemmatizovaný a morfologicky označovaný korpus?**

Možnosti vyhledávání v korpusu:

Výchozí atribut: lemma | sublemma | word

Nabídka výchozího atributu je závislá na konkrétním korpusu, na použité lemmatizaci a značkování.

**Regulární výrazy** ([http://wiki.korpus.cz/doku.php/pojmy:regularni\\_vyrazy](http://wiki.korpus.cz/doku.php/pojmy:regularni_vyrazy))

Konkordance, KWIC

Konkordance představuje všechny doklady (výskyty) hledaného jevu v korpusu spolu s **okolním kontextem**. V praxi se v rámci konkordance rozlišuje **KWIC** (tj. key word in

context), tedy hledané slovo/jev a jeho pravý a levý kontext. Jeden řádek konkordančního seznamu se označuje jako konkordanční řádek.

syn2020 ★

Pokročilý dotaz  | Klávesnice | Předchozí dotazy | Interpretace dotazu

kočka

TIP Na hodnotu tagu lze kliknout s přidržením CTRL pro vyvolání interaktivního nástroje (další tip)

- Upřesnit parametry

Shoda velikosti písmen  ? Povolit regulární výrazy  Výchozí atribut: lemma | sublemma | word

+ Specifikovat kontext

+ Omezit hledání

Hledat

Korpus: syn2020 | Dotaz: kočka (5 637 výskytů) ▶ Promíchat: ✓

Výskytů: 5 637 | i.p.m.: 46,27 (vztaženo k celému korpusu) | ARF: 1 879,81 | Výsledek je promíchaný 1 / 141 ▶▶

Výběr řádků: základní

<input type="checkbox"/>	Λα	neobyčejně dobrácký kocour ( Mirabeau , Gingernut – jako všechny	kočky	měl spoustu jmen ) mi dělal společnost při psaní ,
<input type="checkbox"/>	Λα	„ Počkej ! “ chtělo se Máje vykřiknout . Po	kočce	ale jako by se zem slehla . Mája ji hledala
<input type="checkbox"/>	Λα	co psa nemají . 60 ( Mimochoodem , mezi vlastníky	koček	a lidmi bez koček rozdíl nebyl . ) V roce
<input type="checkbox"/>	Λα	chrámu . Bohoslužbu slova povede František Hladký . Mezinárodní výstava	koček	v hradeckém Černigovu Na Mezinárodní výstavě ušlechtilých koček můžete vidět
<input type="checkbox"/>	Λα	teď kytice chvojí tu v Praze připomíná les . Oběma	kočkám	jsme jednou dopřáli , aby měly kořata : ty jsou
<input type="checkbox"/>	Λα	„ A vysvětlilo by se tak , jak se	kočka	z lodi dostala do opatství . “ Když si všimla
<input type="checkbox"/>	Λα	I v tom , že si dělá starosti o soužití	kočky	a psa . Třebaže ten obrovský apatický kocour jménem Kulička
<input type="checkbox"/>	Λα	to vyjednávání je nebo bylo pro kočku , spíš ta	kočka	se naučí manýrům než tihle mocipáni . A to SKD
<input type="checkbox"/>	Λα	že strašidelný zvuk , který jsme slyšeli , vydávala drobná	kočka	. Byla černá jako rozlitý inkoust a její žluté oči
<input type="checkbox"/>	Λα	na 35,7 miliardy dolarů . Za krmivo pro psy a	kočky	Češi utratili sedm miliard MALOOBCHOD Útraty Čechů za krmivo pro
<input type="checkbox"/>	Λα	a jsou tam vzácné z toho důvodu , že zdravá	kočka	v každé zemi zavraždí ročně v průměru sto ptáků .
<input type="checkbox"/>	Λα	teprve sotva pár měsíců a vypadal spíš jako trochu přerostlá	kočka	, už tehdy tento poněkud výstřední domácí mazlíček některé obyvatele
<input type="checkbox"/>	Λα	. Konečně si dokázala udržet osamělost , tu starou divokou	kočku	, daleko od těla . Začala Satakemu říkat , kocourku

## Zobrazení

KWIC/Věta

[WaG](#) [KonText](#) [SyD](#) [Morfio](#) [KWords](#) [Treq](#) | [Wiki](#) [Support](#) [Biblio](#) | [Zobrazení](#) [Nápověda](#)

**kon text**    Dotaz   Korpusy   Uložit   Konkordance   Filtr   Frekvence   Kolokace   **Zobrazení**   Nápověda

Korpus: [syn2020](#) | Dotaz: [kočka](#) (5 637 výskytů) ▶ Promíchat: ✓

Výskytů: 5 637 | l.p.m.: 46,27 (vztaženo k celému korpusu) | ARF: 1 879,81 | Výsledek je promíchán    1 / 141

Výběr řádků: základní

- Na osobnější úrovni děkuji svým rodičům , kteří mě velice dobrosrdečně poslouchovali větami typu „ Už jsi TO dokončil ? “ Má manželka a já jsme během posledních týdnů redakčních úprav ztratili milovaného a vzácného společníka : náš neobyčejně dobrácký kocour ( Mirabeau , Gingernut – jako všechny **kočky** měl spoustu jmen ) mi dělal společnost při psaní , někdy seděl vedle počítače , někdy na něm .
- Po **kočce** ale jako by se zem slehla .
- Opět zjistili majitelé psů zemřou do roka po infarktu s podstatně menší pravděpodobností než ti , co psa nemají . 60 ( Mimochoodem , mezi vlastníky **koček** a lidmi bez koček rozdíl nebyl . )
- Mezinárodní výstava **koček** v hradeckém Černigovu
- Oběma **kočkám** jsme jednou dopřáli , aby měly kořata : ty jsou teď někde u lidí v okolí .
- „ A vysvětlilo by se tak , jak se **kočka** z lodi dostala do opatství . “

## Korpusová nastavení

(Lemma, POS – part of speech)

Korpusová nastavení pro [syn2020](#)

[Poziční atributy](#)   [Struktury](#)   [Metainformace](#)   [Rozšiřující funkce](#)

Které atributy zobrazit?

	Zobrazovat	Hlavní
word	<input checked="" type="checkbox"/>	<input checked="" type="radio"/>
lc [lowercase word]	<input type="checkbox"/>	<input type="radio"/>
sforma	<input type="checkbox"/>	<input type="radio"/>
lemma	<input checked="" type="checkbox"/>	<input type="radio"/>
lemma_lc [lowercase lemma]	<input type="checkbox"/>	<input type="radio"/>
sublemma	<input type="checkbox"/>	<input type="radio"/>
sublemma_lc [lowercase sublemma]	<input type="checkbox"/>	<input type="radio"/>
tag	<input type="checkbox"/>	<input type="radio"/>
pos [part of speech]	<input checked="" type="checkbox"/>	<input type="radio"/>
case [grammatical case]	<input type="checkbox"/>	<input type="radio"/>



Výskytů: 5 637 | i.p.m.: 46,27 (vztaženo k celému korpusu) | ARF: 1 879,81 | Výsledek je promíchán 1 / 141

Výběr řádků: základní

<input type="checkbox"/>	neobyčejně dobrácký kocour ( Mirabeau , Gingernut – jako všechny	kočky/kočka/N	měl spoustu jmen ) mi dělal společnost při psaní ,
<input type="checkbox"/>	„ Počkej ! “ chtělo se Máje vykřiknout . Po	kočce/kočka/N	ale jako by se zem slehla . Mája ji hledala
<input type="checkbox"/>	co psa nemají . 60 ( Mimochoodem , mezi vlastníky	koček/kočka/N	a lidmi bez koček rozdíl nebyl . ) V roce
<input type="checkbox"/>	chrámu . Bohoslužbu slova povede František Hladký . Mezinárodní výstava	koček/kočka/N	v hradeckém Černigovu Na Mezinárodní výstavě ušlechtilých koček můžete vidět
<input type="checkbox"/>	teď kytice chvojí tu v Praze připomíná les . Oběma	kočkám/kočka/N	jsme jednou dopřáli , aby měly kořata : ty jsou
<input type="checkbox"/>	„ A vysvětlilo by se tak , jak se	kočka/kočka/N	z lodi dostala do opatství . “ Když si všimla
<input type="checkbox"/>	I v tom , že si dělá starosti o soužití	kočky/kočka/N	a psa . Třebaže ten obrovský apatický kocour jménem Kulička

## Metainformace

Korpusová nastavení pro syn2020

Pozíční atributy    Struktury    **Metainformace**    Rozšiřující funkce

<input type="checkbox"/> <#> <input type="checkbox"/> Počet tokenů	<input checked="" type="checkbox"/> <doc> <input type="checkbox"/> Pořadí dokumentu <input checked="" type="checkbox"/> doc.title <input type="checkbox"/> doc.subtitle <input type="checkbox"/> doc.author <input type="checkbox"/> doc.issue <input type="checkbox"/> doc.publisher <input type="checkbox"/> doc.pubplace <input type="checkbox"/> doc.pubyear <input type="checkbox"/> doc.first_published <input type="checkbox"/> doc.translator <input type="checkbox"/> doc.srclang <input type="checkbox"/> doc.authsex <input type="checkbox"/> doc.transsex <input type="checkbox"/> doc.txtype_group <input type="checkbox"/> doc.txtype <input type="checkbox"/> doc.genre_group <input type="checkbox"/> doc.genre	<input type="checkbox"/> <text> <input type="checkbox"/> text.author <input type="checkbox"/> text.section <input type="checkbox"/> text.section_orig <input type="checkbox"/> text.id	<input type="checkbox"/> <p> <input type="checkbox"/> p.id	<input type="checkbox"/> <s> <input type="checkbox"/> s.id
---	--	--	---	---

kon text    Dotaz    Korpusy    Uložit    Konkordance    Filtr    Frekvence    Kolokace    Zobrazení    nápověda

Korpus: syn2020 | Dotaz: kočka (5 637 výskytů) ▶ Promíchat: ✓

Výskytů: 5 637 | i.p.m.: 46,27 (vztaženo k celému korpusu) | ARF: 1 879,81 | Výsledek je promíchán 1 / 141

Výběr řádků: základní

<input type="checkbox"/>	Evropa devatenáctého století	neobyčejně dobrácký kocour ( Mirabeau , Gingernut – jako všechny	kočky/kočka/N	měl spoustu jmen ) mi dělal společnost při psaní ,
<input type="checkbox"/>	Věznění	„ Počkej ! “ chtělo se Máje vykřiknout . Po	kočce/kočka/N	ale jako by se zem slehla . Mája ji hledala
<input type="checkbox"/>	Geniální ps	co psa nemají . 60 ( Mimochoodem , mezi vlastníky	koček/kočka/N	a lidmi bez koček rozdíl nebyl . ) V roce
<input type="checkbox"/>	Deníky Bohemia	chrámu . Bohoslužbu slova povede František Hladký . Mezinárodní výstava	koček/kočka/N	v hradeckém Černigovu Na Mezinárodní výstavě uší
<input type="checkbox"/>	Jsme v nebi	teď kytice chvojí tu v Praze připomíná les . Oběma	kočkám/kočka/N	jsme jednou dopřáli , aby měly kořata : ty jsou
<input type="checkbox"/>	Holubice smrti	„ A vysvětlilo by se tak , jak se	kočka/kočka/N	z lodi dostala do opatství . “ Když si všimla
<input type="checkbox"/>	Záhada mrtvých nohou	I v tom , že si dělá starosti o soužití	kočky/kočka/N	a psa . Třebaže ten obrovský apatický kocour jméno

## Kompletní info o zdrojovém textu:

doc.title	Geniální psi	doc.subtitle	
doc.author	Hare, Brian - Woods, Vanessa	doc.issue	
doc.publisher	Dokořán	doc.pubplace	Praha
doc.pubyear	2016	doc.first_published	2016
doc.translator	Houserová, Jana - Houser, Pavel	doc.srclang	en: angličtina
doc.authsex	X: neuvedeno	doc.transsex	X: neuvedeno
doc.txtype_group	NFC: oborová literatura	doc.txtype	POP: populárně naučná literatura
doc.genre_group	NAT: přírodní vědy	doc.genre	AGR: zemědělství, chovatelství
doc.medium	B: kniha	doc.periodicity	NP: neperiodická publikace
doc.audience	GEN: obecné publikum	doc.isbnissn	978-80-7363-775-0
doc.biblio	Hare, Brian - Woods, Vanessa (2016): Geniální psi. Překlad: Houserová, Jana - Houser, Pavel. Praha: Dokořán.	doc.id	hare_genialnips
text.author		text.section	
text.section_orig		text.id	hare_genialnips:1
p.id	hare_genialnips:1:1074	s.id	hare_genialnips:1:1074:4

Cvičení:

Najděte v korpusu SYN2020 slovní tvar (atribut **word**) *jedle*.

Jak je tvar interpretován lemmatem a tagem?

Najděte v korpusu SYN2020 slovní spojení *příliš jedle, vypadá jedle*.

Jak je tvar interpretován lemmatem a tagem?

přehnané, ale fialová zelenina se zrcadlovým leskem opravdu nevypadá **příliš/příliš/Db-----jedle/jedle/NNFS1-----A-----**, navíc syrový lílek (především mladý a sklizený před

**kon text**    Dotaz Korpusy Uložit Konkordance Filtr Frekvence Kolokace Zobrazení Nápoředa

Korpus: **syn v8** | Dotaz: **vypad.\*;jedle** (8 výskytů) ▶ Promíchat: ✓

Výskytů: **8** | i.p.m.: **0** (vztaženo k celému korpusu) | ARF: **5,1** | Výsledek je promíchán 1 / 1

Výběr řádků: **základní**

<input type="checkbox"/>	vajíčka . Zastavuji se u pultu s uzeninami . Čabajka	<b>vypadá/vypadat/VB-S---3P-AA---I jedle/jedle/NNFS2-----A-----</b>	. A ty pářečky . Dívám se do peněženky .
<input type="checkbox"/>	první jarní květy . A teď mi řekněte , jak	<b>vypadá/vypadat/VB-S---3P-AA---I jedle/jedle/NNFS1-----A-----</b>	Děti z hradecké mateřské školy Klíček se vs
<input type="checkbox"/>	, slouží často jako dekorace . Mangostan Na první pohled	<b>nevypadá/vypadat/VB-S---3P-NA---I jedle/jedle/NNFS1-----A-----</b>	: slupka je tmavá , až do černa . Když
<input type="checkbox"/>	do původní polohy , " přemítá . Na konci roku	<b>vypadala/vypadat/VpFS---3R-AA---I jedle/jedle/NNFS2-----A-----</b>	takhle ... a teď stojí znovu na svém místě
<input type="checkbox"/>	dětech , které ve zdevastované půdě hledají cokoli , co	<b>vypadá/vypadat/VB-S---3P-AA---I jedle/jedle/NNFS1-----A-----</b>	". Potravinové přídělky pro řadové obyvatele
<input type="checkbox"/>	Je libo návod pro inspiraci ? Ve společné skupině dobře	<b>vypadá/vypadat/VB-S---3P-AA---I jedle/jedle/NNFS1-----A-----</b>	ojiněná , borovice wintergold , která na zim
<input type="checkbox"/>	být podle doktorky Lenderové jakákoli zelená dřevina . Nejlépe samozřejmě	<b>vypadá/vypadat/VB-S---3P-AA---I jedle/jedle/NNFS1-----A-----</b>	, ale svůj úkol důstojně plní i smrk nebo bor
<input type="checkbox"/>	často bývají napadeny plísněmi . A tak , i když	<b>vypadají/vypadat/VB-P---3P-AA---I jedle/jedle/NNFP1-----A-----</b>	, mohou být zrádné . Mění se u nich i

Úkol na příště:

Prostudovat [www stránky ÚČNK](http://www.uctnk.cz)

**Umět odpovědět na otázky:**

1. Co je to korpusu?
2. Co je to Český národní korpus?
3. Jaké typy korpusů máme k dispozici?
4. Co to znamená, když řeknu, že korpus má 100 milionů slov?
5. Jak komunikujeme s korpusem (jak jej můžeme využívat pro lingvistickou práci)?
6. Jak můžeme vyhledat v korpusu výskyt slova, jak se se zobrazí v korpusu výskyt slova a co můžeme se zobrazenými výskyty dále dělat?
7. Jak můžeme vyhledat v korpusu všechna slova, která mají společnou vlastnost, že jsou tvary jednoho základního tvaru?
8. Jak můžeme v korpusu vyhledat všechny tvary na rovině gramatické abstrakce (třeba podstatné jména rodu ženského ve 3. pádě, nebo slovesa v přítomném čase v první osobě)?

**A připravit si otázky, na něž byste rádi znali odpověď (souvisí s korpusy!!)**